생체 네트워크와 네트워크를 이용한 단백질 기능 예측

글_이주용 · 미국 NIH 연구원(前 고등과학원 연구원)

20세기 후반, 월드와이드웹 분석을 통해 척도 없는 네트워크(scale-free network)의 존재와 그것의 다양한 일반적인 성질들 – 멱급수 분포(power-law distribution), 두꺼운 꼬리(fat-tail), 허브(hub)노드의 존재 등 – 이 알려지기 시작하였다. 이러한 척도 없는 네트워크의 성질은 다양한 다른 학분 분야의 복잡계 분석에 커다란 영향을 끼쳤고, 그 중에서도 생물학은 가장 많은 영향을 받은 분야 중 하나이다. 2000년대 초반 인간게놈프로젝트의 완성 및 실험 기술의 발전에 힘입어 생물학적 정보의 양은 비약적으로 늘어났고, 네트워크 이론은 이러한 대규모 정보를 처리하고 이해할 수 있는 기반을 제공하였다.

이후, 많은 연구를 통해 월드와이드웹에서 발견된 척도 없는 네트워크의 성질들이 신진대사 네트워크(metabolic network) 및 단백질 상호작용 네트워크(protein-protein interaction network)와 같은 생체 네트워크에서도 존재함이 알려졌다. 서로 전혀 상관이 없는 것처럼 보이는 월드와이드웹과 생체 네트워크 사이에 존재하는 유사성의 이유는 무엇일까? 척도 없는 네트워크의 가장 큰 특징 중 하나는 네트워크에 존재하는 노드들이 공격을 받더라도 전체 네트워크의 연결성은 큰 영향을 받지 않는다는 점이다. 반면에 무작위로 연결된 네트워크의 경우 소수의 노드들이 공격을 받는다면 전체 네트워크가 서로 연결되지 않은 작은 네트워크로 단시간 내에 분절됨이 잘 알려져 있다. 이러한 성질은 생명체가 외부의 공격에 최소한으로 영향을 받고 항상성을 유지할 수 있도록 진화하였음을 시사한다. 또한, 신진대사 네트워크에서 비슷한 위상적 특성을 가지는 노드들은 다른 종에서도 비슷한 역할을 수행하는 것으로 알려져 있다. 이러한 발견은 생체 네트워크가 생명체의 근본적인 구성 원리에 대한 해답을 찾는데 중요한 역할을 할 수 있다는 것을 보여준다.

2000년대에 가장 활발히 연구되었던 생체 네트워크는 신진대사 네트워크와 단백질 상호작용 네트워크였지만 최근에는 그 생물학적 대상이 점차 다양해지고 있는 추세이다. 최근 많은 관심을 끌고 있는 네트워크로는 대표적으로 질병 네트워크를 들 수 있다. 질병 네트워크는 동일한 환자에게서 같이 발견되는 질병들을 연결한 네트워크에서 시작해서, 질병-약 네트워크, 질병-유전체 네트워크 등으로 그 개념이 확대되고 있으며, 실용적인 측면에서는 신약개발에도 도움을 줄 것으로 기대되고 있다. 기존 신약 개발의 일반적인 방법은 특정 질병에 관련된 단백질을 찾은 후, 그 단백질의 기능을 저해하는 물질을 찾는 것이었다. 그러나 최근 들어 이러한 방법의 한계점이 점차 드러나고 있다. 일반적으로 생명체 안에서 어떠한 단백질의 활동이 저해되면, 생명체는 그 단

백질의 기능을 대체할 수 있는 다른 경로들을 찾아내게 된다는 점이다. 이러한 문제를 해결하기 위해, 하나의 특정한 단백질에만 선택적으로 작용하는 물질을 찾는 대신에 특정 질병에 관련된 단백질들을 네트워크상에서 찾은 후, 찾아낸 다수의 단백질들과 동시에 상호작용할 수 있는 물질을 찾는 방법이 많은 관심을 받고 있다.

네트워크 모델의 측면에서도, 기존의 한 가지 종류의 노드와 연결을 가진 네트워크 모델에서 확장된, 다양한 종류의 노드와 연결을 가진 모델에 대한 연구가 활발히 진행되고 있다. 대표적인 예로 최근 정하웅 교수 연구단에서 연구한 단백질-결합체 이분 네트워크(protein-complex bipartite network)를 들 수 있다. 이 네트워크에서 각각의 단백질은 그 단백질이 참여하는 단백질 결합체와 연결을 이루고 있으며, 단백질과 단백질은 결합체를 통해서 연결된다. 이러한 네트워크는 한 가지 종류의 노드와 연결을 가진 단백질 상호작용 네트워크에 비해 더욱 자세한 정보를 제공해 줄 수 있으며, 각 노드들이 가지는 연결수의 분포도 기존에 잘 알려진 멱급수가 아닌 지수분포(exponential distribution)를 따르는 것이 밝혀졌다. 이러한 결과는 동일한 생물학적 계를 기술하더라도 사용된 네트워크 모델에 따라 다른 통찰을 얻을 수 있음을 보여준다. 위에서 언급된 질병-약 네트워크 및 질병-유전체 네트워크도 이와 같은 이분 네트워크 모델로 기술될 수 있다. 일반적인 네트워크에 비해 이분 네트워크의 성질과 특징은 아직 밝혀져야 할 부분이 많이 남아있다.

생체 네트워크를 이용한 대표적인 응용 분야 중 하나는 상호작용 네트워크를 이용한 단백질 기능 예측이다. 다양한 서열 분석 실험 기법의 발전으로 인해 많은 생물들의 유전자 서열 정보는 알려져 있지만, 아직까지 그 기능에 대해서는 알려지지 않은 부분이 많이 남아 있다. 가장 많이 연구된 생물체인 효모균(yeast)의 경우에도 전체 단백질 중 1/4의 기능은 아직 밝혀지지 않은 상태이다. 만일 이미 알려진 정보를 바탕으로 알려지지 않은 단백질의 기능을 정확하게 예측할 수 있다면 실험에 필요한 많은 시간과 자원을 아낄 수 있을 것이다. 단백질 상호작용 네트워크를 바탕으로 단백질의 기능을 예측하는 방법을 만드는 것은 현재 생물정보학 분야에서 가장 활발하게 연구되고 있는 주제 중 하나이다.

상호작용 네트워크를 이용한 단백질 기능 예측 방법들은 크게 두 가지 범주 즉, 연결되어 있는 이웃 노드의 정보를 바탕으로 하는 방법과 네트워크에서 서로 밀접히 연결되어 있는 무리(module/cluster)의 정보를 이용하는 방법으로 분류할 수 있다. 이웃 노드의 정보를 이용하는 방법들은 다시 단순히 이웃한 노드들이 가지고 있는 기능의 빈도(frequency)에 의존하는 방법과 마코프 모델(Markov model)을 이용한 방법으로 나눌 수 있다. 빈도에 의존하는 방법에 따르면 기능이 알려지지 않은 특정 노드는 이웃 노드들에서 가장 많이 발견되는 n개의 기능을 가지고 있을 것이라 추정된다. 마코프 모델의 경우, 임의의 노드가 임의의 기능 f를 가질 확률은 이웃한 노드들 중에 기능 f를 가진 노드의 개수와 가지지 않은 노드의 개수를 변수로 가지는 로지스틱 함수에 의해 결정된다. 이 로지스틱 함수의 형태를 결정하는 매개변수들은 기능이 알려진 네트워크 상의 다른 노드들의 정보를 최대우도추정법(maximum likelihood estimation)을 이용해서 결정된

다. 이러한 이웃한 노드의 정보를 이용하는 방법의 장점은 계산 속도가 매우 빠르고 비교적 쉽게 계산을 수행할 수 있다는 것이다. 그러나 실험데이터의 부족으로 인하여 노드들의 기능이 많이 알려지지 않았거나, 노드들이 비교적 적은 연결수를 가지고 있는 네트워크의 경우 예측의 정확도는 제한적일 수밖에 없다.

일반적으로 생체네트워크는 서로 더 밀접하게 연결되어 있는 작은 무리들로 나뉠 수 있음이 알려져 있다. 이러한 무리들은 기능적으로 연관되어 있으며, 물리적으로도 서로 단백질 복합체를 이룰 가능성이 높은 것으로 알려져 있다. 이 같은 성질들을 바탕으로 네트워크상에서 같은 무리 안에 속해 있는 노드들은 서로 비슷한 기능을 공유할 것이라는 가정 하에 단백질의 기능을 예측하는 다양한 방법들이 제안되고 있다. 노드들의 무리를 이용한 방법은 복잡한 네트워크를 좀 더 이해하기 쉬운 형태로 바꾸어 주는 장점이 있다. 그러나 무리 안에 속해 있는 모든 노드들이 같은 기능을 가진다는 가정 때문에 많은 거짓 양성 예측(false positive prediction) 결과를 주는 단점을 가지고 있다. 최근 수행된 비교 연구에 따르면 일반적으로 이웃 노드의 정보를 이용하는 방법이 무리 정보를 이용하는 방법보다 좋은 결과를 주는 것으로 확인되었다.

최근 고등과학원의 이주영 교수 연구단에서는 노드의 네트워크상의 위상적 정보, 이웃한 노드들의 정보, 무리 정보를 모두 고려하여 기존에 제안된 방법들보다 더 정확하게 단백질의 기능을 예측할 수 있는 방법을 제안하였다. 이 방법은 네트워크에서 얻을 수 있는 다양한 정보들을 기계학습(machine learning) 방법을 이용하여 통합함으로써 통계적 예측 모델을 만드는 것이다. 새로운 방법은 외부에서 추가 정보가 주어지지 않더라도 기존의 방법들보다 더 정확한 예측 결과를 얻을 수 있다는 장점이 있다. 또한, 이 방법은 주어진 네트워크의 구조에서 얻을 수 있는 정보에만 의존하기 때문에 단백질 기능 예측뿐만 아니라 다른 생체 네트워크에도 쉽게 적용이 가능하다. 따라서 앞으로 다양한 생체 네트워크 분석에 도움을 줄 것으로 기대한다.

참고 문헌

1. Barabási, A. L., & Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. Nature Reviews Genetics, 5(2), 101–113.

2. Albert, R., Jeong, H., & Barabási, A. L. (2000). Error and attack tolerance of complex networks. Nature, 406(6794), 378–382.

3. Goh, K. I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., & Barabási, A. L. (2007). The human disease network. Proceedings of the National Academy of Sciences, 104(21), 8685–8690.

4. Hopkins, A. L. (2008). Network pharmacology: the next paradigm in drug discovery. Nature chemical biology, 4(11), 682–690.

5. Sharan, R., Ulitsky, I., & Shamir, R. (2007). Network-based prediction of protein

function. Molecular systems biology, 3(1).
6. Lee, S. H., Kim, P. J., & Jeong, H. (2011). Global organization of protein complexome in the yeast Saccharomyces cerevisiae. BMC systems biology, 5(1), 126.
7. Song, J., & Singh, M. (2009). How and when should interactome-derived clusters be used to predict functional modules and protein function?. Bioinformatics, 25(23), 3143-3150.
8. Lee, J., Gross, S., & Lee, J. (2013). Hidden information revealed by optimal community structure from a protein-complex bipartite network improves protein function prediction. PLOS One, in press.

이주용

서울대학교 화학과에 학부생으로 입학하여 박사학위까지 취득한 뒤 연구원으로 근무하였다. 그 다음은 고등과학원 계산과학부에서 연구원으로, 현재는 미국국립보건원(National Institutes of Health)의 연구원으로 재직중이다.